DOCUMENT RESUME

ED 043 678 TM 000 154

AUTHOR Myers, Charles T.
TITLE A Covariance Approach to Item Analysis.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Sep 70
NOTE 13p.

EDRS PRICE EDRS Price MF-$0.25 HC-$0.75
DESCRIPTORS Analysis of Variance, Correlation, *Item Analysis, *Reliability, Scores, *Statistical Analysis, Test Construction, *Test Reliability

ABSTRACT

This paper brings together a variety of item-analysis techniques into a coherent system. The system is based on classical test theory, the theorems that can be derived from the equation, $X = T + E$. The system extends from techniques for analyzing parts of an item separately to techniques for relating items to total test score, to sub-scores, to external criterion scores, and to test-retest situations. The system is both mathematically simple and basic and may serve to bring test theory and test practice closer together. (Author)

TEST DEVELOPMENT REPORT
TDR-70-1 SEPTEMBER 1970

A Covariance Approach to Item Analysis

Charles T. Myers
Test Development Division, ETS

A Covariance Approach to Item Analysis

Charles T. Myers
Educational Testing Service

Abstract

This paper brings together a variety of item-analysis techniques into a coherent system. The system is based on classical test theory, the theorems that can be derived from the equation, $X = T + E$. The system extends from techniques for analyzing parts of an item separately to techniques for relating items to total test score, to sub-scores, to external criterion scores, and to test-retest situations. The system is both mathematically simple and basic and may serve to bring test theory and test practice closer together.

A Covariance Approach to Item Analysis

Charles T. Myers

Educational Testing Service

Introduction

A variety of different item analysis techniques have been developed to serve a variety of purposes in test construction and test theory. A number of these item analysis techniques provide two indices for each item, usually a "difficulty" index and an index showing a relationship between score on the item and score on the test of which the item is a part, an index of item-test "homogeneity." There are three different types of correlation coefficient that are commonly used for this purpose: the biserial, the point-biserial, and the *phi* coefficients. With this much variety available to the test developer and the test theorist (not to mention other possibilities such as the use of item characteristic curves), it does not seem necessary to advance a new alternative. Actually, it is the purpose of this paper to discuss some of the advantages of an old but rarely used approach and to indicate how it can be extended to some new techniques and purposes. This approach appears to have the advantage of simplicity so that it is easy to understand, to compute, and to criticize. It seems to have a minimum of assumptions and it is closely related to some of the basic concepts of test theory. Its principal distinction is that it uses covariances rather than correlations for its indices of homogeneity.

With a square matrix of variances and covariances for any set of variables, the sum of all the variances and covariances equals the variance

of the sum of variables. If the variables are the items in a test, the sum of all the item variances and inter-item covariances equals the variance of the test, the sum of a row or column equals an item-test covariance, and the sum of the item-test covariances also equals the test variance. This is true whether the items are scored dichotomously zero and one or for any other scoring system, but with zero and one scoring there an some interesting simplifications. For one thing, if the mean score of an item (the proportion passing in this case) is known then the item variance is fixed, the variance is dependent on the mean; which is, of course, not true for the general case of score distributions.

A correlation matrix may be thought of as a covariance matrix of variables with standardized scores, and a common procedure for computing a correlation matrix is first to find the covariance matrix and then divide each row and column by the appropriate standard deviation. The principal advantage of correlations as compared with covariances is to assist in the interpretation of these statistics, just as standardized test scores assist in score interpretation. The principal disadvantage of correlations is that it is more difficult to interpret the meaning of sums of correlation coefficients than it is to interpret the meaning of sums of covariances. Also, there is still controversy over the appropriate technique for standardizing item-test correlations--some favor biserials and others favor point-biserials. Since items that are scored zero and one are standardized as to range and since item variance is fixed by item difficulty, standardization may not be so useful for items and it may be more useful in item analysis to use covariances rather than correlations. It should be noted that an item-test

covariance may be divided by the standard deviation of the test to produce a statistic that seems to share some of the advantages of covariances and some of the advantages of correlations. In this paper this statistic will be called the item effectiveness index and will be symbolized by e_{it}.

Procedures

In classical test theory (the theory derived from the equation $X = T + E$), the three principal statistical characteristics of a test are its mean, its standard deviation (or its variance), and its reliability. This covariance approach to item analysis is so closely related to these aspects of classical test theory that it may be appropriate to call it classical item analysis. This system of item analysis provides three indices for each item. These indices are each respectively related to the test mean, the test standard deviation, and the test reliability. In each case the test statistic is obtained merely by summing the item indices. This system also provides techniques for evaluating parts of the test and even parts of the items--for splitting the test atom. Finally, this system includes a technique for evaluating test and item validity when criterion data are available. Although in its simplest form this system assumes that items are scored either zero or one, it may easily be extended to formula-score tests.

The first moment of a test score distribution is its mean and the first statistic in this classical covariance approach to item analysis is the item mean. The item mean is the proportion passing the item,

symbolized p_i . The sum of the proportions passing all the items in a test equals the test mean. The item variance is determined by the item mean and equals the item mean times one minus the item mean, $s_i^2 = p_i(1 - p_i)$ (Horst, 1966).

The second moment of a test score distribution is its variance and the second statistic in this item analysis system is the item effectiveness index, e_{it} . This index is computed by finding the item-test covariance and dividing it by the standard deviation of the test. The sum of item effectiveness indices for all the items in the test equals the test standard deviation. One value that is gained by using the item effectiveness index instead of the item-test covariance is that it facilitates comparisons between items taken from tests of different lengths. This index is the same as the item-analysis index that was called the "reliability index" in Gulliksen's <u>Theory of Mental Tests</u> (1950) and symbolized $r_{xg}s_g$. It is obvious that a correlation coefficient multiplied by the standard deviation of one variable is equal to the covariance divided by the standard deviation of the other variable.

Many item analysis systems provide only two indices for each item. The two indices that have been described for this system may be used for most of the purposes that item analysis has been used for, both for test production and for elementary classical test theory. However, the covariance approach is compatible with a number of other statistics that are logical extensions of the system. The first of these relates item analysis to test reliability through Woodbury's (1951) concept of the standard length of a test. The reliability discussed here is the Kuder-Richardson (1937) formula 20 reliability. This third item coefficient has been called the "length" of the item and has been symbolized by k_i (Myers, 1961).

The standard length of a test, as defined by Woodbury, is the number of items required for the test to have a reliability of .50. If the standard length is taken as a unit, then the length of an item is the fraction of that unit that is represented by a single item. If Kuder-Richardson formula 20 reliability is understood to be the ratio of true variance to observed variance, it implies that the true variance of an item is equal to the average of the covariances of that item with all the other items in the test. If that average is subtracted from the item variance, the remainder is understood to be error variance by this definition. The sum of these remainders is equal to the variance of errors of measurement for the test. The length of an item is computed by dividing the average covariance of the item by the remainder or error variance. The sum of the item lengths is equal to the length of the test in standard length units. The reliability of the test is easily computed by dividing the test length (in standard length units) by one plus that length.

There is another possible use for the average inter-item covariance statistic. It has often been found difficult to interpret an item analysis of sub-scores or part scores in a test. Typically these sub-scores are fairly highly positively correlated and the distinctions between them are subtle. Many item analysis homogeneity coefficients include an element, sometimes called a spurious element, produced by the perfect correlation of the item with itself. This element makes the interpretation of the subtle differences between sub-scores very difficult. When the average inter-item covariance is computed as it was in the previous paragraph, the item variance is not included in the average. Thus this statistic should clarify the

analysis of tests into homogeneous sub-scores.

The covariance approach to item analysis lends itself conveniently not only to the analysis of tests into sub-scores, but also to the analysis of a single item into its parts. As anyone knows who has found a miskeyed item in an item analysis, it is possible to do distracter analyses as well as item analyses. The homogeneity index for a distracter is usually negative. In the covariance analysis, the sum of the effectiveness indices for all the responses to an item is always equal to zero; therefore the sum of the indices for all distracters equals minus one times the index for the correct answer, that is, if no one omits the item. If some omit the item, that response can also be analyzed. Thus, the standard deviation of a test equals minus one times the sum of the effectiveness indices for all the responses other than the correct responses. Sums of separate categories of these responses may be of interest. For example, a test speededness index may be computed from the sum of the effectiveness indices for all the responses of not reaching an item.

Gulliksen (1950) has shown how item indices may be used to study the validity of items when a relevant criterion score is available. The same procedure might also be used with scores on a retest or parallel test administered after a learning interval. Using the retest as the criterion should offer some new insights into the difficult problem of measuring gain rather than the traditional static measurement of position.

Although this already appears to be an extensive and comprehensive system of item analysis, it is quite possible that other uses and extensions of this system can easily be developed.

Discussion

The covariance system of item analysis does not appear to have been widely used, yet it appears to have many worthwhile advantages. It can be applied to most of the common uses of item analysis in the art of producing tests. It is simpler to compute than many other systems and it can be applied to test production by any man who can add and who has a pencil and pad of paper. It involves the most simple and direct relationships between item statistics and score distribution statistics. This simplicity also provides an element of mathematical elegance that might appeal to even a sophisticated test theorist. Good test production requires highly sophisticated subject matter competence on the part of the test assembler. There has often been a difficulty in communication between such persons and mathematical test theorists. Perhaps the greatest value of the covariance approach to item analysis is that it may bring these two branches of expertise more effectively together.

References

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Horst, P. Psychological measurement and prediction. Belmont, Calif.: Wadsworth Pub. Co., 1966.

Kuder, G. F. & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Myers, C. T. A note on the standard length of a test. Psychometrika, 1961, 26, 443-446.

Woodbury, M. On the standard length of a test. Psychometrika, 1951, 16, 103-106.

Appendix

Summary of Classical Item Analysis

Where M_i is the mean score on item i, and

p_i is the proportion passing item i, then

$$M_i = p_i \ . \tag{1}$$

Where s_i^2 is the variance of scores on item i, then

$$s_i^2 = p_i(1 - p_i) \ . \tag{2}$$

Where M_t is the mean score on test t, and

n is the number of items in test t, then

$$M_t = \sum_{i=1}^{n} p_i \ . \tag{3}$$

Where c_{it} is the covariance of item i and test t, then

$$e_{it} = c_{it}/s_t \ . \tag{4}$$

Where e_{it} is called the "item effectiveness" index of item i in test t, and

c_{ij} is the covariance of item i and item j, then

$$c_{it} = s_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} c_{ij} \ , \tag{5}$$

and

$$s_t^2 = \sum_{i=1}^{n} c_{it} = s_t \sum_{i=1}^{n} e_{it} \ , \tag{6 and 7}$$

so that

$$s_t = \sum_{i=1}^{n} e_{it} \ . \tag{8}$$

Where r_{bis} is the item-test biserial correlation, and

$r_{p.bis}$ is the item-test point-biserial correlation, and

y_i is the ordinate of the normal curve at the point that cuts off a proportion equal to p_i, then

$$e_{it} = r_{bis} y_i = r_{p.bis} s_i \, . \tag{9 and 10}$$

$$\bar{c}_i = (c_{it} - s_i^2)/(n - 1) \, , \tag{11}$$

and

$$\bar{c} = (s_t^2 - \sum_{i=1}^{n} s_i^2)/(n^2 - n) = \sum_{i=1}^{n} \bar{c}_i/n \, . \tag{12 and 13}$$

Where k_i is the "length" of item i in standard length units, then

$$k_i = \bar{c}_i/(\bar{s}_i^2 - \bar{c}) \, , \tag{14}$$

and where r_{tt} is the test reliability as defined by Kuder-Richardson formula 20,

$$r_{tt} = \sum_{i=1}^{n} k_i/(1 + \sum_{i=1}^{n} k_i) \, , \tag{15}$$

and where u_i is defined as $s_i^2 - \bar{c}_i$, then

$$r_{tt} = n^2\bar{c}/(n^2\bar{c} + n\bar{u}) \, . \tag{16}$$

Where d represents distracters in a 5-choice item and if no one omits the item, then

$$e_{it} = (-1)(\sum_{d=1}^{4} e_d) \, . \tag{17}$$

Where M_{t_f} is the mean score of test t computed by the formula:

$$\text{score} = \text{right} - \frac{1}{4}\,\text{wrong, and}$$

s_{t_f} is the standard deviation of formula scores, and

p_{+_i} is the proportion answering correctly, and

p_{-_i} is the proportion answering incorrectly, then

$$M_{t_f} = \sum_{i=1}^{n} p_{+_i} - \frac{1}{4} \sum_{i=1}^{n} p_{-_i} \, , \qquad (18)$$

and

$$s_{t_f} = \sum_{i=1}^{n} e_{+_{it}} - \frac{1}{4} \sum_{i=1}^{n} e_{-_{it}} \, . \qquad (19)$$